

A Honeypot for the Exploration of Spammers' Behavior

by Guido Schryen

Spam has become one of the most annoying and costly phenomenon on the Internet. Valid e-mail addresses are among the most valuable resources of spammers, but little is known about the methods by which spammers collect and harvest addresses. Spammers' capabilities and interest in carefully directed, consumer-oriented marketing have not yet been explored. Gaining insight into spammers' ways of obtaining and misusing e-mail addresses is useful in many ways; *e.g.*, for assessing the effectiveness of techniques that obscure addresses and the usefulness and necessity of hiding e-mail addresses on the Internet. This paper presents a spam honeypot project in progress that addresses these issues by systematically placing e-mail addresses on the Internet and analyzing received e-mails.

The Threat

Spam is generally recognized as an increasingly disturbing and costly issue for electronic business and Internet traffic. Companies, non-profit organizations, and individuals receive this type of e-mail to such an extent that the issue has certainly gone beyond that which is merely "annoying." Symantec reports that, in scanning 100 billion e-mails, the percentage of spam e-mails reached 69% in January 2005 but decreased to 60% in May. [1] MessageLabs announced that the average global ratio of spam was nearly 70% in May 2005, although the sample of e-mails inspected was much smaller, comprising some one million per day. [2] The content of spammers' e-mails covers a broad range of topics:

- Offering or advertising general goods and services, such as devices, investigative services, clothing, and makeup (21% of all e-mails categorized as spam)
- Containing references or offerings related to money, the stock market, or other financial "opportunities" (19%)
- Containing or referring to products or services intended for persons above the age of 18 (10%)
- Offering or advertising health-related products and services (13%) [1]

The increased payload of networks and e-mail servers and the demand on employees' time and attention are not the only harmful effects of spam e-mails. Fraudulent messages; *e.g.*, e-mails that appear to be from a well-known company but are not—also known as "brand spoofing" or "phishing" e-mails—are often used to trick users into revealing personal information, such as e-mail addresses, financial information, and passwords (7%). Furthermore, viruses, worms, and Trojan horses (opening backdoors for botnets using the infected computer as a spam client) are distributed over the Internet. The total economic damage caused by spam e-mails is estimated at several billion dollars. [3]

This central economic aspect has motivated anti-spam activities embracing many facets: national laws and international regulations (about which Hintz [4] provides a good overview); organizational provisions, including abuse systems (*e.g.*, <http://spam.abuse.net/>) and lists of suspicious domains and IP numbers; and technical solutions that mainly apply blocking, filtering, or authenticating mechanisms. [5] Statistics and e-mail users' daily experience show that the spam problem is far from being solved, and it is only by applying technical anti-spam that the collapse of our Internet e-mail system has been prevented.

Implementing honeypots and honeynets has emerged as a solution [6, 7], along with these mainstream efforts to analyze spammers' behaviour or to even attack them. The honeypot presented here contributes to this field by setting up a technical environment that analyzes where spammers get their e-mail addresses and how they exploit them—or if they simply use any harvested e-mail address. (A more detailed presentation of the honeypot project can be found in Schryen, [8]).

Motivation and Goals

Valid e-mail addresses are among the most valuable resources of spammers, and identifying address sources and the procedures used by spammers to exploit them is crucial to preventing spammers from getting addresses and misusing them. It is widely known that, besides generating addresses with brute-force mechanisms, spammers get valid e-mail addresses by harvesting the Internet or, illegally, from organizations. Some Address Obscuring



Techniques (AOTs) that restrict the availability and usability of e-mail addresses have been proposed: As early as 1997, Hall [9] described e-mail channels, and in 2003, Ioannidis [10] presented a policy for encapsulating single-purpose addresses. Many users also use temporary addresses and dispose of them when they feel that the spam quotient has become too high.

Gaining insight into spammers' ways of obtaining and misusing e-mail addresses is useful in many ways:

- Assessing the effectiveness of AOTs and input for their improvement
- Identifying spammers to lead to their prosecution
- Assessing the usefulness and necessity of hiding e-mail addresses on the Internet
- Discovering specific marketing and addressing activities

The last item, above, focuses on the quality of e-mail addresses. Spammers are known to collect as many valid e-mail addresses as possible, but little is known about spammers' capabilities and interest in carefully directed, consumer-oriented marketing. A taxonomy of quality for e-mail addresses is shown in Figure 1.

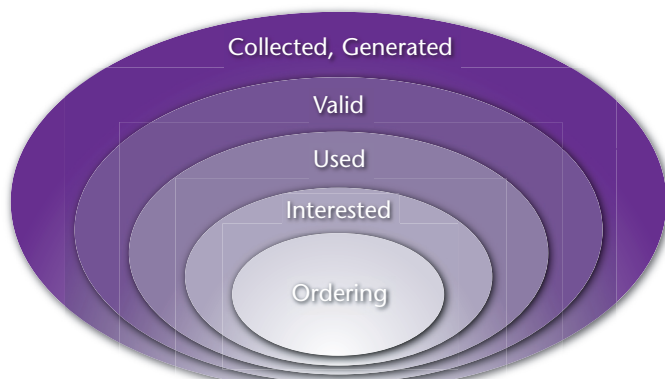


Figure 1: Taxonomy of E-mail Addresses

The inner ellipses are more valuable for spammers than the outer ones because of losses caused by non-selective advertising. Only a portion of collected or generated e-mail addresses are valid ones, *i.e.*, e-mails addressed to non-valid ones are refused by the addressee's host because these mailboxes do not exist. Valid ones can be divided into addresses actually in use and those that are no longer accessed and thus useless for spammers. A way to distinguish between the two is provided by an "opt-out" option included in some spam e-mails; however, when this option is used incautiously by the spam recipient, it indicates that the address is in use. Spammers will go even further and adopt physical marketing strategies using knowledge about consumer-specific interests and behaviour; *e.g.*, an Internet user actively participating in a German discussion group that focuses on medical products is presumably interested in offers of medical products in the German language. The innermost ellipse contains e-mail addresses of users who buy products and thus from whom the spammer profits.

The goal of the honeypot is to (1) penetrate spammers' behavior in harvesting e-mail addresses from Internet services, such as newsgroups and the Web, and (2) to discover the extent to which spammers have already shifted from simply employing e-mail addresses in use towards acquiring addresses of users likely to be interested in specific marketing.

Conceptual Framework

To cover a broad range of locations that are attractive to spammers for harvesting e-mail addresses, it is necessary to inspect many Internet services. Integrated into this honeypot are newsletters and mailing lists, Web pages, Web chats, chats, and the Usenet in which e-mail addresses are placed. There are many more ways in which spammers can get e-mail addresses [11] that have not yet been covered. This is simply caused by the limited resources of the project, which is currently not funded.

To detect linguistic and regional particularities, each medium is divided into those that are oriented to the German language and those that are US based. This furnishes a second, desirable dimension in that it renders the study readily extensible to other languages

and regions. To inspect spammers' behavior regarding specific marketing activities, a third dimension of the survey focuses on the topic of the Internet service. For example, Web pages and newsletters and mailing lists are divided into those ruled by an individual, a discussion board, a greeting-card service, *etc.*, in which the topics are grouped by types of administration, content, connection, context, and commerce. (For a complete list of topics, see Schryen, [8]) It should be noted that topics are service specific. Figure 2 shows the classification of Internet locations as used in the empirical study. Each type of location is represented by a cube, each cube contains three locations (a location is a specific Web site or a specific newsletter), each location gets four addresses (de-, com-, net-, and org-address), and for each cube 12 e-mail addresses must be reserved. This procedure makes it possible to detect if the top-level domain of an e-mail address is relevant. So far, German and US newsletters and mailing lists and Web pages have been addressed, *i.e.* the number of e-mail addresses placed for getting harvested is almost $2 \times 2 \times 36 \times 12$, which is 1728. Of course, no e-mail address must be seeded more than once.

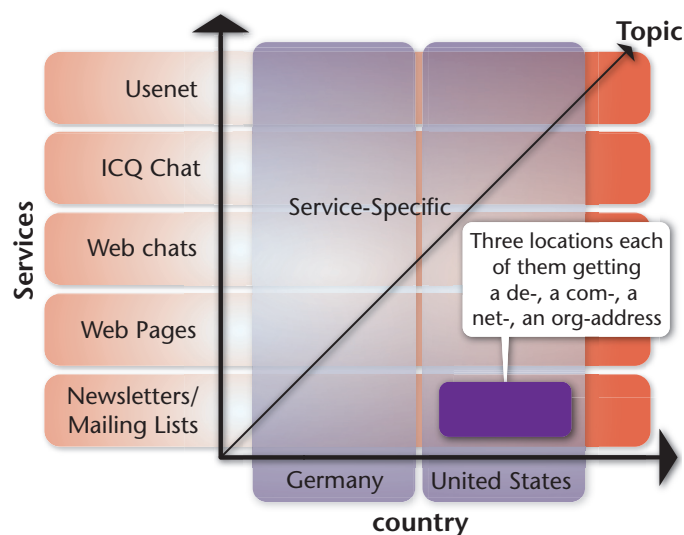


Figure 2: Classification of Internet Locations

Implementation

A mail server has been set up, charlie.winfor.rwth-aachen.de, and three domains have been reserved, wforasp.com, wforasp.net, and wforasp.org, to cover the e-mail addresses of four top-level domains. All e-mails addressed to these domains are directed to this mail server. As thousands of e-mail addresses had to be created, they were automatically generated by a random generator for the user part of the addresses. To prevent e-mail addresses from being guessed or generated with brute-force attacks, it is necessary to define them randomly and to give them an appropriate number of characters. An example of an e-mail address created this way is wasp10208@wforasp.com. The Internet locations serving as lures were chosen manually, just as the placement of the e-mail addresses had to be done manually. As soon as an e-mail address is spread, its location and activation date is stored.

All incoming e-mails are classified into regular e-mails (ham e-mails), such as regular newsletters or the like that contain comments from users of discussion forums, and spam e-mails. This procedure is currently mainly executed

by humans but supported by a mail parser written in Hypertext Preprocessor (PHP), which uses an increasing white list containing pairs of recipient-addresses, Internet Protocol (IP) entries: each time a host was manually assessed as qualified to send an e-mail to the recipient address, its IP number was linked to this e-mail address and stored in the white list. A second task of the mail parser is to decompose each incoming e-mail—all entries of the header and the content are analyzed, as is the Multipurpose Internet Mail Extensions (MIME) structure of the body. (A detailed description of the relational data model on which the procedure is based is beyond the scope of this paper.) Next, the e-mails' elements are stored in the Structured Query Language MySQL database broken down into spam and ham e-mails. The database is intended to be used by data-mining tools and (simpler) statistical analyzers. Figure 3 provides a survey of the implementation infrastructure.

First Empirical Results

In total, 15,178 ham e-mails and 8,189 spam e-mails have been recorded by our mail server. Because of the very early stage of the project, the results presented here are preliminary; however, some facts are worth mentioning:

- No spam has been sent to addresses that were used for subscribing German newsletters/mailling lists.
- Only a few spam e-mails have been received by way of US newsletter/mailling list subscription. The few are all due to administration topics.
- Not surprisingly, many more spam e-mails arise from placements on web pages. Interestingly, German web pages were responsible for only a third of the number of spam e-mails that are due to US web pages. Net-addresses seem to be of greater interest to spammers than de- and org-addresses independently of any country; on US web sites com-addresses have been even more used by spammers.

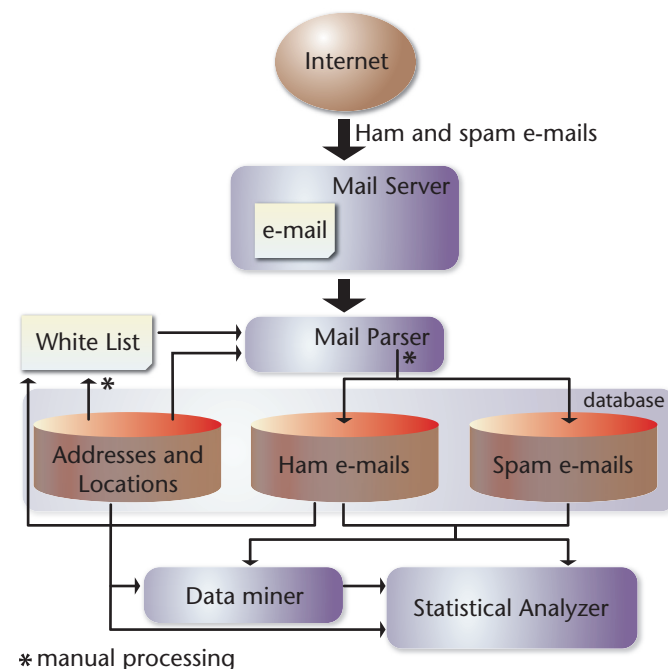


Figure 3: Infrastructure of the E-mail Honeypot Environment

Summary and Outlook

Spammers are known to collect as many valid e-mail addresses as possible, but little is known about spammers' capabilities and interest in carefully directed, consumer-oriented marketing. Gaining insight into spammers' ways of obtaining and misusing e-mail addresses is useful for

- Assessing the effectiveness of AOTs
- As input for the improvement of AOTs
- For identifying spammers leading to their prosecution
- For assessing the usefulness and necessity of hiding e-mail addresses on the Internet
- For discovering specific marketing and addressing activities

This article sketches a honeypot to penetrate spammers' behavior in harvesting e-mail addresses from Internet services, such as newsgroups and the Web, and in discovering the extent to which spammers have already shifted from simply employing e-mail addresses already in use toward acquiring addresses of users likely to be interested in specific marketing offers. The honeypot's conceptual framework classifies Internet locations as used in the empirical study using three dimensions: Internet services, (*e.g.*, the Usenet, Web pages, newsletters); service-specific topics such as education, infotainment, auctions; and countries. Each location gets four addresses (*de-*, *com-*, *net-*, and *org-*addresses), which permits the researcher to detect if the top-level domain of an e-mail address is relevant for spammers. When e-mails arrive at the honeypot's mail server, they are classified into spam and ham e-mails (regular e-mails), decomposed by a parser, and stored in a database that is intended to be used by data-mining tools and (simpler) statistical analyzers. Preliminary results of the honeypot study are presented, which show that no spam has been sent to addresses that were used for subscription to German newsletters and mailing lists, that only a few spam e-mails have been received due to US newsletter and mailing-list subscriptions, that many more spam e-mails arise from placements on Web pages, and that *net-* as well as *com-*addresses seem to be of particular interest to spammers.

The project is at an early stage. More services and countries remain to be integrated, more data must be collected for more reliable results, and a time-series analysis must be applied. Another avenue that needs to be explored is the functional; *i.e.*, the application of data-mining procedures and statistical procedures aiming at detecting differences between spam and ham e-mails. These results can be used to improve spam filters.

The experiences gained from the prototypic honeypot implementation can be used to develop a general blueprint for further honeypots that explore spammers' behaviour and the effectiveness of AOTs. Depending on funding, software tools will be included to enable a semi-automated setup and utilization of future honeypots. ■

References

- [1] Symantec, Spam statistics, <http://www.symantec.com/region/de/PressCenter/spam.html> [Accessed 04/01/05].
- [2] MessageLabs, Email Threats, <http://www.messagelabs.com/emailthreats/default.asp> [Accessed 04/01/05].
- [3] OECD, Background Paper For The OECD Workshop On Spam, 2003.
- [4] Hintz T., Opt-In vs. Opt-Out Legislation, <http://notebook.ifas.ufl.edu/spam/Legislation.htm> [Accessed 04/01/05].
- [5] Schryen, G, Effektivität von Lösungsansätzen zur Bekämpfung von Spam, *Wirtschaftsinformatik* 46 (2004) 4, pp. 281–288. (English version is not published but is available from the author.)
- [6] The Honeynet Project. <http://honeynet.org>. [Accessed 04/01/05].
- [7] Project Honey Pot. <http://www.projecthoneypot.org>. [Accessed 04/01/05].
- [8] Schryen, G., An e-mail honeypot addressing spammers' behavior in collecting and applying addresses. Proceedings of the 6th IEEE Information Assurance Workshop, West Point, pp. 37–41.
- [9] Hall, R., Channels: Avoiding Unwanted Electronic Mail. Proceedings DIMACS Symposium on Network Threats DIMACS, 1996.
- [10] Ionnadis, J, Fighting Spam by Encapsulating Policy in Email Addresses. Network and Distributed System Security Symposium (NDSS'03), 2003.
- [11] Raz, U., How do spammers harvest email addresses? <http://www.private.org.il/harvest.html> [Accessed 04/01/05]

Acknowledgements

The setup of the honeypot was strongly supported by Reimar Hoven. The classification of incoming e-mails had to be performed manually, to which task Stephan Hoppe dedicated much time. Many thanks are also due to Jan Herstell and to Katrin Ungeheuer for proofreading.

About the Author

Guido Schryen

Mr. Guido Schryen graduated from the RWTH Aachen University (Germany), where he earned a Masters' degrees in Computer Science and in Operations Research. He received his PhD from the Faculty of Business Administration and Economics of RWTH Aachen University where he now holds a postdoctoral position. His current research activities focus on Internet security and anti-spam measures. He may be reached at schryen@winfor.rwth-aachen.de.